My PhD began as an effort to understand and solve some of the practical problems that arose during my four and a half years as an engineer at Google. My initial goal was to understand how to propagate uncertainty in large problems for which classical Bayesian techniques (namely Markov Chain Monte Carlo) are prohibitively expensive. This problem eventually led to a more general investigation of sensitivity analysis in both Bayesian and frequentist statistics. My work is theoretical at times, but I have always maintained the goal of providing practical and easy-to-understand tools for scientists who are trying to rigorously solve difficult data problems. Having developed a suite of practical tools and theoretical frameworks, I now believe that my research will be best served by applying these ideas to practical problems in close collaboration with practicing scientists. For this reason, I am applying to become an Alvarez Fellow.

An Example Problem from Astronomy.

During the initial years of my PhD I collaborated on an LBNL project that can provide a useful practical context for describing potential applications of my PhD work. Astronomers build catalogues of celestial objects from sky surveys consisting of very large numbers of telescopic images that tile the night sky. Images are typically taken in a multiple different color bands, may overlap, and are blurred by an unknown "point spread function" (PSF). It would be desirable to combine images from multiple surveys which may be taken at different resolutions. From the pixels in these images, astronomers wish to infer the location of celestial objects as well as some key properties, such as color, brightness, and, in the case of galaxies, shape and orientation. One way to do so is to posit a parameterized generative model, from which the unknown true location and properties of celestial objects combine with random noise to produce the observed images. Given this generative model, one tries to infer what the unknown true catalogue might have been using the images and Bayesian statistical techniques.

In this problem, uncertainty and correlation abounds. Consider two dim stars which are close together on the scale of the pixelized image. It can be difficult to ascertain whether the image is of two distinct stars, or a single oblong galaxy. The relative probabilities of each depends on the unknown PSF, about which there may also be some uncertainty. One may wish to borrow strength between multiple images, and take into account how the multiple images' different resolutions affects the amount of information they give. Ideally, we would take all these tangled uncertainties into account when stating our final probabilistic belief about the identity of the object(s).

Linear Response Covariances.

On small problems, one would typically quantify all this uncertainty with Markov Chain Monte Carlo (MCMC). However, with the vast amount of data contained in a sky survey, MCMC is far too computationally expensive. Researchers instead turn to "mean field variational Bayes" (VB), an optimzation-based

approximation to the full Bayesian solution.[1] Howvever, the very reason that MFVB is tractacble is because it assumes that there is no correlation between disparate aspects of the model. For example, for the purpose of inferring the identity of a star, the PSF is treated as known and fixed, not uncertain. By making the problem computationally tractable, we have to discard some of the correlations we wish to account for in the first place. Indeed, MFVB is notorious for producing unusuably small estimates of posterior uncertainty.

The first section of my thesis addresses this problem with a technique which we call "linear response variational Bayes" (LRVB). By considering the sensitivity of the MFVB approximation to perturbations of the objective, one can recover an approximation to the full Bayesian posterior. We showed that, in a wide set of typical models (taken from the Stan Examples datasets), LRVB allows Bayesian posterior covariances to be accurately approximated orders of magnitude faser than MCMC.

The LRVB covariance estimate has a crucial computational property common to all of my PhD research, which is that the required sensitivity can be computed without ever re-solving the initial optimization problem. What is needed, instead, is the solution to a system of linear equations involving the derivatives of the MFVB objective function at the optimum. In the case of the sky survey, this means LRVB requires the solution of only one optimization problem—the original MFVB fit. Since most of the inferred parameters don't affect one another (e.g., the value of PSF on one night does not affect the classicificaiton of a star on a different night), the linear system is sparse. Furthermore, though even storing the full covariance matrix would be prohibitively expensive, individual covariance estimates can be queried using iterative methods such as the conjugate gradient algorithm. Finally, the required derivatives, which would be tedious and error-prone to compute by hand, can be quickly and faultlessly computed automatically by modern automatic differentiation.

Cross Validation.

Modern automatic differentiation and scalable, iterative linear solvers allow the automatic computation of the sensitivity needed to compute the LRVB covariance estimates of the previous section. It turns out that many other key statistical tasks can also be accurately approximated by linear sensitivty analysis. In this section,

[1] The frequentist "expectation-maximization", or EM algorithm, can in fact be understood as a MFVB algorithm, and is so included in my discussion here.